International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 6, No. 1, (2022), pp. 111-118

Graduation Prediction Application using Naive Bayes Algorithm in Sukabumi Muhammadiyah University

Indra Griha Tofik Isa[1], Indra Satriadi[2]*

[1,2]Informatics Management, Politeknik Negeri Sriwijaya, Indonesia

Corresponding E-mail: indra_isa_mi@polsri.ac.id

Abstract

The use of data mining has become a trend in data processing because of the availability of large amounts of data and the increasing need to convert this data into useful information and knowledge. Besides as a tool in extracting data, data mining is also used as decision support, both in the commercial and non-commercial fields. Of the many algorithms used in data mining, one of them is the Naïve Bayes Algorithm, which in this algorithm is one of the methods in Probabilistic Reasoning which aims to classify data in certain classes. The study conducted by applying the stages of the Naïve Bayes Algorithm in application design to predict student graduation on time based on the parameters contained in the Enrolment Student (PMB), Grade Point Average (IP) and the Finance Department. The data processed were 436 datasets from the Informatics Engineering Study Program, Muhammadiyah University of Sukabumi. The system design uses UML modeling with implementation using the PHP programming language and MySQL database. The results of the study are in the form of graduation prediction applications with an accuracy rate of 78%

Keywords: *Data Mining, Classification, Naive Bayes*

1. Introduction

Data mining is how to translate the phenomena that occur, which unconsciously are very useful information and knowledge. Currently, anything uses information technology media in processing data, so a lot of data is resulted and even wasted only to be used as the output of a system process. Whereas these data can be used as useful information for the stakeholders. Of course, to extract the data required a method, one of which is Data Mining. The algorithms used in data mining also vary, including [5]: C4.5, K Means, Support Vector Machine, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, CART. Muhammadiyah University of Sukabumi is the largest university in Sukabumi City with 7 faculties and 20 study programs, one of which is the Informatics Engineering Study Program (IT Study Program). The number of students owned by the IT Study Program in the 2019/2020 Academic Year amounted to 314 students [1]. There are several parameters that become a reference in assessing the success of the Study Program, one of the indicators is the graduation rate on time, which means that a student completes his studies within a period of 7-8 semesters. When viewed from the trend in the number of graduates on time, taking the difference between final year students and the number of graduates, the number of graduates from 2017 to 2020 tends to fluctuate. In 2017 the on-time graduation rate was 87.30%, in 2018 it was 81.16%, in 2019 it was 70.91%, in 2020 it was 89.06%. Graphically, it can be seen in Figure 1.

ISSN : 2580-7250

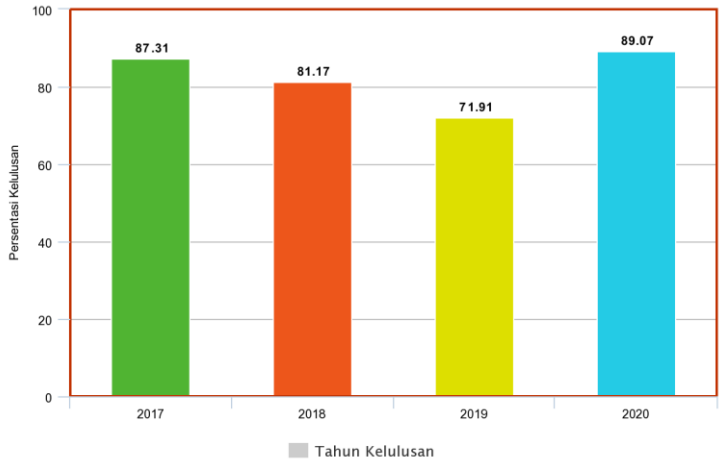

Figure 1. "On-Time" Graduation Rate of Informatics Engineering Study Program, Year of 2017 to 2020

Many factors can affect the graduation rate on time. Student data can be used as a reference in determining the graduation rate on time. However, a method is needed in extracting information from student data, one of the method is using the Data Mining Algorithm, that is Nave Bayes. The correlation in this study is how to extract student information based on NIM, Student Name, GPA, First Choice of Major, Second Choice of Major, Percentage of Payment History (level of payment accuracy) to predict graduation on time. The results of these data are used as parameters in designing applications for predicting student graduation on time.

2. Reseach Methodology

2.1. Data Mining

Data mining is the mining or discovery of new information by looking for certain patterns or rules from a very large amount of data [2]. Data mining is also defined as a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning [3] to extract and identify useful information and related knowledge from various large databases [4].

The process of extracting added value in the form of knowledge / Knowledge Discovery in Database (KDD) [5] in Data Mining is (1) Selection; (2) Pre-Processing / Cleaning; (3) Transformation; (4) Data Mining; (5) Interpretation. Data Mining is divided into several groups based on the tasks that can be done [6], those are (1) Description; (2) Estimation; (3) Prediction; (4) Classification; (5) Clustering; (6) Association.

2.2. Naive Bayes Algorithm

One of the algorithms used in data mining is the Naïve Bayes algorithm, which is part of the classification technique [7]. Naive Bayes was developed by Thomas Bayes by combining statistical and probability techniques [8]. Naive Bayes predicts opportunities based on data or previous experience where it is assumed that the conditions between attributes are independent [9]. Naive Bayes classification assumes that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes. The formulation of Naive Bayes as follow [10]:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (1)$$

Description:

X : Data with the unknown class
H : Hypothesis data X that is a specific class
P(H|X) : Probability of hypothesis H based on condition X (Posteriori Probability)

P(H) : Hypothesis probability H (Prior Probability)
P(X|H) : Probability of X based on the conditions on the hypothesis H
P(X) : Probability of X

2.3. Selection of used data

This is the initial stage, where this research finds some student information taken from Enrolment Student data (PMB), Student Academic Data (SIAK) and Student Financial Data for the 2013-2016 PMB period and 2017-2020 student graduate data. The total data used are 436 datasets. The resulting data is raw data that will be processed in the next stage.

2.4. Reduction the unused data

From the raw data in the previous stage, the next step is sorting the data that will be used, from this stage the data used are NIM, Student Name, GPA, First Choice of Study Program, Second Choice of Study Program, Year of Graduation, Percentage of Late Payments.

2.5. Implementation of Data Mining

This stage is part of the implementation of data mining from the designed data. Application Development using Microsoft Visual Studio 2012 with MySQL database. The raw data is inputted to be processed by the Nave Bayes algorithm which has been translated into a programming language, which produces an output in the form of recommendations. The system design uses an object-oriented approach with modeling Unified Modeling Language.

3. Result and Discussion

3.1. Selection of Used Data

The initial data used in this study is an excel dataset contained in the Enrolment Student (PMB), Academic Information System (SIAK) and Student Financial Data. The data processed is data on new student admissions for the Academic Years of 2013/2014, 2014/2015, 2015/2016 and 2016/2017. The processed data is 60% of the PMB data, with a total of 436 datasets. The data taken is student data that is active for 7-8 semesters, for PMB student data with Drop Out status not included in the dataset.

Table 1. PMB Data Period of 2013-2017

No	Name	Year Admission	PIL_1	PIL2
1.	Siti Risalti	2015	TI	SASING
2	Juanda	2015	TI	ADBIS
3	Vebryan	2016	TI	TS
4	Maria Mulfah	2016	TI	TS
5	Adi Fadilah	2013	TI	TS
…	…	…	…	…

The SIAK data taken is an 8th semester student based on the Year the student became UMMI Informatics Engineering student. The number of student datasets corresponds to the number of PMB datasets, which are 436 datasets. Data taken from SIAK is NIM, Student Name, GPA, Year of Graduation (on time estimate of 8 semesters). Meanwhile, from financial data, payment data is obtained, by calculating the number of late payment history divided by the total payment period. The dataset obtained is in the form of NIM, Student Name, Percentage of Late Payments.

Table 2. Financial Data Periode of 2013-2017

No	Name	Payment Method	% Payment
1.	Siti Risalti	SMT	0%
2	Juanda	SMT	10%

No	Name	Payment Method	% Payment
3	Vebryan	SMT	0%
4	Maria Mulfah	SMT	0%
5	Adi Fadilah	SMT	20%
…	…	…	…

3.2. Reduction the Unused Data

From PMB, SIAK and Finance data, not all data were used, the parameters involved were NIM, Student Name, First Choice of Study Program, Second Choice of Study Program, Year of Graduation and Late Payment Rate (in %). At this stage, the dataset is tested using the Naïve Bayes algorithm to see the pattern of the data. All attributes will contribute to decision making, with the same weight of attributes being important and each attribute independent of each other. If given k independent attributes, the probability value can be given as follows:

$$P(x_1 \ldots x_k|C) = P(x_1|C)x \ \ldots \ xP(x_k|C) \quad (2)$$

The initial stage of the Naïve Bayes calculation is to collect training data from student data (PMB, SIAK, Finance). The determining variables used in classifying students' "on-time" graduation data are:

a) First Choice of Study Program (Pil1_Jur)
   It is the first choice variable for new students in determining the study program they choose
b) Second Choice of Study Program (Pil2_Jur)
   It is the second choice variable for new students in determining the study program they choose
c) Year of Graduation (Tahun Lulus)
   Student graduation year, with an estimated 8 semesters (4 years) of new student enrollment.
d) Payment Method (Cara Bayar)
   Is a variable method of payment for lectures, categorized into 3 categories, namely Monthly (BLN), Semester (SMT) and Annual (THN)
e) % Payment (% Bayar)
   Is a variable for the level of student late payments, taken from the history of the intensity of student late payments divided by the intensity of payments per semester
f) Ontime (Tepat Waktu)
   It is the difference between the Year of Graduation and the Year of Enrollment. Punctuality is categorized into 2 categories namely ONTIME (TEPAT) and NOT ONTIME (TIDAK TEPAT)

The sample data can be seen on table 3 as follows:

Tabel 3. Student Dataset

No	Student Name	Tahun Daftar	Pil1_ Jur	Pil2_ Jur	Thn Lulus	IPK	Cara Bayar	% Bayar	Tepat Waktu
1	Siti Risalti	2015	TI	SASING	2019	3.45	SMT	0%	TEPAT
2	Juanda	2015	TI	ADBIS	2019	3.11	SMT	10%	TEPAT
3	Vebryan	2016	TI	TS	2020	3.70	SMT	0%	TEPAT
4	Maria Mulfah	2016	TI	TS	2020	3.26	SMT	0%	TEPAT
5	Adi Fadilah	2013	TI	TS	2019	3.01	SMT	20%	TIDAK TEPAT
6	Sarah Septian	2015	TS	TI	2020	2.75	BLN	10%	TIDAK TEPAT
7	Windri A	2014	TI	ADPUB	2018	3.04	BLN	0%	TEPAT
8	Raditya	2013	TI	AKT	2017	3.26	THN	0%	TEPAT
9	Jihan Fideyna	2016	TI	AKT	2020	2.86	THN	10%	TEPAT
10	Feri	2015	TI	SASING	2019	2.93	THN	0%	TIDAK

No	Student Name	Tahun Daftar	Pil1_ Jur	Pil2_ Jur	Thn Lulus	IPK	Cara Bayar	% Bayar	Tepat Waktu
	Andriansyah								TEPAT
…	…	…	…	…	…	…	…	…	…

3.3. Implementation of Data Mining

In calculating the Naïve Bayes algorithm, it is necessary to translate the table 2 dataset into binary form (1 and 0), so it is concluded that there are 4 variables, namely **Pilihan1 Jurusan, IPK, % Bayar dan Tepat Waktu**. Each variable has 2 values and the class variable also has 2 values, as explained below:

Variable: **Pilihan 1 Jurusan**

a. IT : 1
b. Non IT : 0

Variable: **IPK**

a. IPK >= 3 : 1
b. IPK < 3 : 0

Variable: **% Bayar**

a. If 0% : 1
b. If > 0%: 0

Variable: **Tepat Waktu**

a. TEPAT : 1
b. TIDAK TEPAT : 0

The results of the dataset conversion can be seen in Table 4 below:

Table 4. Dataset Convertion Result

Student Name	Pil1_ Jur	IPK	% Bayar	Tepat Waktu
Siti Risalti	1	1	1	1
Juanda	1	1	0	1
Vebryan	1	1	1	1
Maria Mulfah	1	1	1	1
Adi Fadilah	1	1	0	0
Sarah Septian	0	0	0	0
Windri A	1	1	1	1
Raditya	1	1	1	1
Jihan Fideyna	1	0	0	1
Feri Andriansyah	1	0	1	0
…	…	…	…	…

There are two types of probabilities that need to be calculated from the dataset for the table above, namely Class Probabilities (Tepat Waktu) and Conditional Probabilities. Calculation of Class Probabilities (On Time) is done by: Terdapat dua jenis peluang yang perlu dihitung dari dataset untuk tabel di atas, yakni Class Probabilities (Tepat Waktu) dan Conditional Probabilities. Perhitungan Class Probabilities (Tepat Waktu) dilakukan dengan:

a. P(Tepat Waktu=1) = Amount(Tepat Waktu=1) / (Amount(Tepat Waktu=1) + Amount (Tepat Waktu=0))
b. P(Tepat Waktu=0) = Amount(Tepat Waktu=0) / (Amount(Tepat Waktu=1)+ Amount(Tepat Waktu=0))

Then resulted:

a. P(Tepat Waktu=1) = 7 / (3+7) = **0.7**
b. P(Tepat Waktu=1) = 3 / (3+7) = **0.3**

Next is to determine the probability of each of these variables, for example for the GPA variable, then the probability rule based on the Naïve Bayes formula is one of them:

$$P(\text{IPK} \geq 3 | \text{Tepat Waktu}) = \frac{\sum IPK \geq 3 \text{ and } Tepat\ Waktu}{\sum Tepat\ waktu} \quad (3)$$

Likewise with other variables, those calculations are needed to determine the probability of each variable.

3.4. System Design and Implementation

In designing the system, the logic flow or flowchart of the application is shown in Figure 2 below:

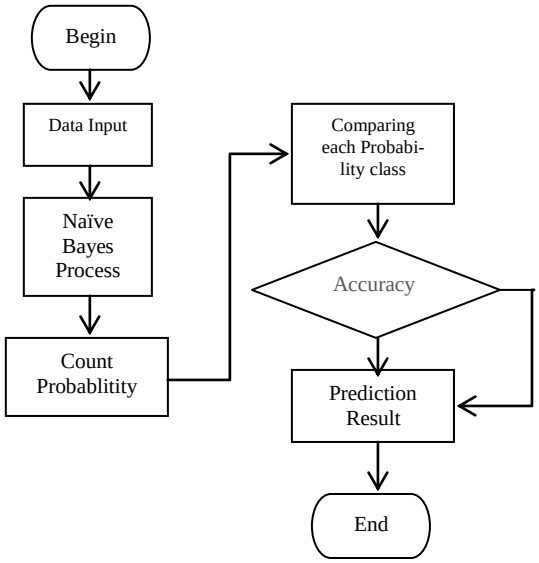

Figure 2. Proposed System Flowchart

In the flowchart above, it begins with data input with predetermined variables, then the Naïve Bayes calculation process is carried out, by calculating the number of probabilities and comparisons of each class from the previously inputted data variables [11]. So that the final result appears predictions of graduation from the inputted student variables.

The system design is carried out by modeling the Unified Modeling Language (UML), with the Use Case diagram in Figure 3 and the class diagram in Figure 3:

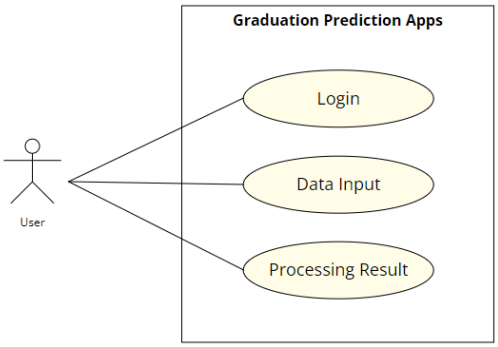

Figure 3. Use Case Diagram of Graduation Prediction Application of Informatics Engineering UMMI

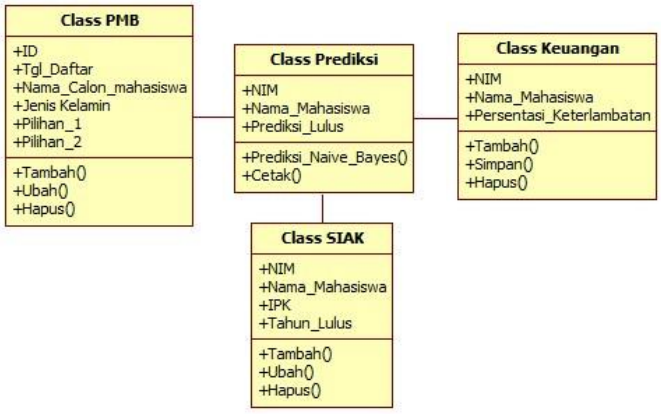

Figure 4. Class Diagram of Graduation Prediction Application of Informatics Engineering UMMI

3.5. Database Design

The following is the design of the database that has been defined by 4 tables, namely PMB table, SIAK table, finance (Keuangan) table, and prediction (Prediksi) table:

Table 5. PMB Table Structure

Field	Data Type	Length
ID	Varchar	20
NIM	Varchar	20
Nama_Mhs	Varchar	50
Tgl_Daftar	Date/Time	Short
Pilihan_1	Varchar	20
Pilihan_2	Varchar	20

Table 6. SIAK Table Structure

Field	Data Type	Length
NIM	Varchar	20
Nama_Mhs	Varchar	50
IPK	Double	
Tahun_Lulus	Integer	

Table 7. Keuangan Table Structure

Field	Data Type	Length
NIM	Varchar	20
Nama_Mhs	Varchar	50
Cara_Bayar	Varchar	20
Persentasi_ Bayar	Integer	

Table 8. Prediksi Table Structure

Field	Data Type	Length
NIM	Varchar	20
Nama_Mhs	Varchar	50
Prediksi_ Lulus	Varchar	20

3.6. System Testing

To measure the quality of the system that developed, a test was carried out by measuring the level of accuracy through inputting test data of 53 student data, from the results of the accuracy calculation there were 12 data that were not match with the real data of student graduation, while the rest where there were 41 data had conformity with real data of student graduation.

From the data above, a percentage calculation for accuracy is carried out by dividing the confirmity data according to the total amount of data entered. So it is calculated (41 confirmity data / 53 input data) x 100% resulting in an **accuracy of 78%.**

5. Conclusion

Based on the research results, it can be concluded that (1) The data mining modeling implemented in the prediction application uses classification modeling; (2) The naive bayes classification modeling implemented in the graduation application has an accuracy rate of 78% based on the test data entered into the application, so that it becomes a recommendation in a graduation prediction based on the parameters of the first choice of Study Program, IPK, and Payment. The suggestions for further application development

are the need for larger data sets to increase the level of accuracy of the implemented modeling.

References

[1] I. G. T. Isa, "Implementasi Pendekatan Kerangka Kerja NIST 800-34 dalam Perancangan Disaster Recovery Plan pada Sistem Informasi Akademik Universitas Implementasi Pendekatan Kerangka Kerja NIST 800-34 dalam Perancangan Disaster Recovery Plan pada Sistem Informasi Akadem," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 15, no. 2, pp. 103–113, 2020, doi: 10.30872/jim.v15i2.3724.

[2] C. C. Aggarwal, "An Introduction to Data Classification," in *Data Classification: Algorithms and Applications*, C. C. Aggarwal, Ed. New York, USA: CRC Press, 2014, pp. 1–31.

[3] A. Smola and S. V. N. Vishwanathan, *Introduction to Machine Learning*. 2014.

[4] D. A. A. AlHammadi and M. S. Aksoy, "Data Mining in Higher Education," *Period. Eng. Nat. Sci.*, vol. 1, no. 2, pp. 1–4, 2013, doi: 10.21533/pen.v1i2.17.

[5] A. Suad A. and B. Wesam S., "Review of data preprocessing techniques in data mining.pdf," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: doi=jeasci.2017.4102.4107.

[6] L. W. Santoso and Yulia, "The Analysis of Student Performance Using Data Mining," *Adv. Intell. Syst. Comput.*, vol. 924, no. June, pp. 559–573, 2019, doi: 10.1007/978-981-13-6861-5_48.

[7] M. Sabransyah, Y. N. Nasution, and F. D. T. Amijaya, "Aplikasi Metode Naive Bayes dalam Prediksi Risiko Penyakit Jantung," *J. EKSPONENSIAL*, vol. 8, no. 2, pp. 111–118, 2017.

[8] I. G. T. Isa, "Aplikasi Asesmen Calon Debitur menggunakan Naive Bayes di Koperasi Mitra Sejahtera SMK Negeri 1 Kota Sukabumi," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 1, pp. 31–39, 2021, doi: 10.32736/sisfokom.v10i1.1013.

[9] Bustami, "Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[10] M. K. Anam, B. N. Pikir, and M. B. Firdaus, "Penerapan Na ̈ive Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen danPemeritah," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[11] J. D. Novaković, A. Veljović, S. S. Ilić, Ž. Papić, and T. Milica, "Evaluation of Classification Models in Machine Learning," *Theory Appl. Math. Comput. Sci.*, vol. 7, no. 1, p. Pages: 39 – 46, 2017, [Online]. Available: https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158.